AD_____

AWARD NUMBER DAMD17-94-J-4383

TITLE: Developing and Implementing the AJCC Prognostic System for Breast Cancer

PRINCIPAL INVESTIGATOR: Philip H. Goodman, M.D., M.S.

CONTRACTING ORGANIZATION: The University of Nevada
Reno, Nevada 89557

REPORT DATE: August 1997

TYPE OF REPORT: Annual

PREPARED FOR: Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for public release;
distribution unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	August 1997	Annual (1 Aug 96 - 31 Jul 97)

4. TITLE AND SUBTITLE
Developing and Implementing the AJCC Prognostic System for Breast Cancer

5. FUNDING NUMBERS
DAMD17-94-J-4383

6. AUTHOR(S)
Philip H. Goodman, M.D., M.S.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The University of Nevada
Reno, Nevada 89557

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, MD 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

14. SUBJECT TERMS
breast cancer

15. NUMBER OF PAGES
28

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

__X__ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and use of Laboratory Animals of the Institute of Laboratory Resources, national Research Council (NIH Publication No. 86-23, Revised 1985).

____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

PI - Signature Date

Table of Contents

ABSTRACT

Accurate survival prediction is important for women with breast cancer because a woman's expected survival determines her therapy, provides her with vital outcome information, and is one of the main selection criteria for entry into new therapy clinical trials. For almost forty years breast cancer outcome prediction has been based on the TNM staging system. This system it is relatively inaccurate, its accuracy continues to decline as screening increases the early detection of breast cancer, and its accuracy cannot be significantly improved. The objective of this research program is to replace the TNM staging system with a computer-based clinical decision support system that provides the most accurate survival predictions possible for women with breast cancer.

Methods

Three goals must be achieved in order to meet the objective of a useful, accurate clinical decision support system for breast cancer. They are: developing the most powerful and efficient statistical techniques, training of the prediction system with breast cancer outcome data, and clinical implementation and validation of the trained system. We have developed a multi-time-interval time-to-event artificial neural network statistical method. We have also developed a gaussian-bernoulli mixture model for binary and continuous missing data. A finite mixture model is being created to group patients by outcome. A measure of discriminative accuracy is being created to assess the accuracy of the artificial neural networks' predictions. We are using several data sets, including the American College of Surgeons' National Cancer Data Base, the National Cancer Institute's Surveillance, Epidemiology and End Results Program, Mayo Clinic data sets, and Duke University data sets that include p53 and HER-2/neu. We are implementing a graphical user interface that makes it easy to enter data and to understand the results. The system will be available for use in desktop and hand-held computers.

Results

The results of the system have been excellent. We have shown that the artificial neural network prognostic system is significantly more accurate in predicting which patients will survive five years than the TNM staging system, principle components analysis, classification and regression trees, and logistic regression. The artificial neural network is also better at ten year survival prediction. We have shown that for early stage breast cancer, the stage usually detected by mammography, the TNM staging system does no better than flipping a coin in predicting who will survive five and ten years. The artificial neural network-based clinical decision support system, with the prognostic factors p53 and HER-2/neu, provides very accurate survival predictions. We have begun to make therapy-specific predictions. We have also shown that the gaussian-bernoulli mixture model is very efficient at dealing with binary and continuous missing data. Our assessment of the finite mixture model and the measure of discriminative accuracy is in progress.

Conclusions

We are reporting work-in-progress, additional tasks must be completed. We will be including additional molecular-genetic prognostic factors to increase our predictive power. We continue to refine the therapy-specific predictions. We will validating the system on other breast cancer data sets and we will implementing a clinical demonstration project to fine tune the system for clinical use.

Introduction

The goal of this project is a computer-based prognostic system for breast cancer that: is significantly more accurate than the TNM staging system, predicts survival over time based on therapy, and presents its predictions in a manner that physicians can understand. This project can be viewed as consisting of three components: (1) Data analysis and prognostic factor evaluation, (2) developing the prognostic model, and (3) implementing a clinically useful system, i.e., breast cancer prognostic factors, the artificial neural network statistical model, and the clinician user interface.

The first year of research was characterized by work on the artificial neural network statistical model and related statistical models, specifically tasks 2.1.2 (artificial neural network generating survival curves), 2.1.3 (determining the accuracy of the survival curves), 1.03, 2.1.4 (comparing the accuracy of the artificial neural network to other statistical models), 2.2 (implementing an effective solution for missing data in training and performance), and 2.3 (dealing with censored data).

In addition, during the first year we started work related to data analysis and prognostic factors including 1.02 and 1.08.1 (recurrence as an endpoint), 1.04.2 (creating a taxonomy of prognostic factors in breast cancer), 1.04.3 (writing a book on prognostic factors in breast cancer, in preparation), 1.06.3 (determining minimum data set size), 1.11 (examining physician breast cancer survival estimates). We also began work on 3.1 (the code) and 3.2 (the physician interface).

Also during the first year we added three tasks, (1) a comparison of the two main American cancer data bases, namely, the Surveillance, Epidemiology, and End Results and the National Cancer Data Base data bases. (2) An examination of the issue of what to do when confronted with cases not lost completely at random and competing risks. (3) Computerization of the TNM staging system.

The second year of research was characterized by the continuation of work begun in the first year and by data analysis and prognostic factor development, specifically tasks 1.01 and 2.1.1 (extending the survival endpoint from five to ten years), 1.02 and 1.03 (see first year), 1.05 (the identification of high risk node negative women), 1.06 (clinical trials), and 1.07 (therapy). Work continued on 2.1.2 (artificial neural network generating survival curves), 2.1.3 (creating a new method for assessing prediction accuracy), 1.03 and 2.1.4 (model comparisons). Work was completed on the computerization of the TNM staging system and the comparison of the two national cancer data bases.

The third year of research was characterized by extending our work to new areas, specifically tasks 1.04 (new prognostic factors) and 1.08 (recurrence) and by beginning the implementation and testing of the system, specifically tasks 3.2 (physician interface) and 3.3 (demonstration project).

Detailed Report by Task

1.04) New prognostic factors in breast cancer.

We have described what we believe to be the correct approach to the evaluation and use of breast cancer prognostic factors. It involves determining the type of prognostic factor (for example, natural history, therapy-dependent, or post-therapy) and, using an appropriate statistical model, assessing its predictive accuracy. (Burke HB, Hoang A, Iglehart JD, Marks JR. Predicting response to adjuvant and radiation therapy in early stage breast cancer. Cancer, in press.)

We have shown that grade was not a very powerful prognostic factor in breast cancer. (Burke HB. Henson DE. Histologic grade as a prognostic factor in breast carcinoma. Cancer 1997;80:1703-1705.)

1.04.1) New prognostic factors

Mammographic early detection of breast cancer is reducing the usefulness of the TNM staging system because most tumors detected by mammography are small and few women have involved lymph nodes or distant metastases. Providing an accurate prognosis in early-detected breast cancer is a critical problem. Can new molecular-genetic prognostic factors take over the predictive burden from the TNM in these women? The answer is probably yes. Using a data set obtained from Duke University (courtesy of Drs. Iglehart and Marks) of 230 women with early detected breast cancer, i.e., small tumors and four or fewer involved lymph nodes, that contained, in addition to the TNM variables, age, estrogen and progesterone receptor status, histology, p53, and erbB-2 we have shown that prognostic accuracy at five and ten years was in the range of .75 - .85 (ROC, area under the receiver operating characteristic). (Burke HB, Hoang A, Iglehart JD, Marks JR. Predicting response to adjuvant and radiation therapy in early stage breast cancer. Cancer, in press.) This is a very encouraging result.

1.04.2) Create a taxonomy of prognostic factors in breast cancer.

Completed, in last year's report.

1.04.3) Prognostic Factors in Breast Cancer book.

Work is ongoing. Burke HB, Henson DE. Prognostic Factors and Systems in Cancer. Kluwer Academic Publishers Inc., in preparation.

1.05) High risk node negative women

Based on our initial results with the new prognostic factors contained in the Duke data set (1.04.1) we believe that some of these factors will be very useful in identifying high risk node negative women. This work is ongoing.

1.08) Extending the analysis of prognostic factors beyond the discovery of the disease.

1.08.1) Recurrence analysis.

Completed, in last year's report.

1.08.2) The important role of time in prognostic factor research.

There are two kinds of time related to prognostic factors. The first is the value of the factor in predicting a future outcome. For example, the accuracy of a factor in predicting five or ten year survival. The second type of time is the predictive value of a factor collected over time. In other words, does the predictive value of the factor change over historical time so that, for example, tumor size is less predictive for women today than it was for women 20 years ago. In collaboration with investigators in Finland we have shown the prognostic value of a factor changes over both types of time. (Lundin M, Lundin J, Burke HB, Toikkanen S, Liisa P, Heikki J. The role of time in breast cancer outcome prediction. Submitted for publication.)

3.2) Physician interface.

It is very important that physicians find the new prognostic system easy to use and useful. To this end we have implemented a Windows interface in C++. We have spent a great deal of time with clinicians (oncologists, pathologists, surgeons, radiation oncologists, and others) adapting it to their needs. This work is ongoing.

3.3) Demonstration project

We have created demonstration projects with Duke University Medical Center and the Mayo Clinic. This has taken a great deal of our time but it has also been very helpful. The results of the demonstration projects will be ready by the end of the grant.

Conclusions

The third year of research was characterized by the more practical aspects of the project. We describe a correct approach to the evaluation and use of breast cancer prognostic factors. We have shown that the new molecular genetic prognostic factors are useful in early detected disease. We better understand the role of time in breast cancer prognostic factors and prognosis. We have been working closely with clinicians on the system interface and on its use. In summary, the research is going very well. We believe that we will be able to successfully meet our goal of providing a computer-based prognostic system that is more accurate than the TNM staging system and that is easy to use and understand within the four year time frame of this grant. In addition, we have created several new systems that we believe will advance the domain of cancer prognosis, e.g., artificial neural network survival-over-multi-interval-time models, an effective missing data method for training and performance, and a new approach to the assessment of prediction accuracy.

Publications and Presentations Related to this Grant During The Third Funding Year

Books

Burke HB, Henson DE. Prognostic Factors and Systems in Oncology. Kluwer Academic Publishers, in preparation.

Burke HB (ed). Artificial Neural Networks in Medicine. Kluwer Academic Publishers, in preparation.

Journals

Rosen DB, Burke HB. Applying a gaussian-bernoulli mixture model network to binary and continuous missing data in medicine. Sixth International Workshop on Artificial Intelligence and Statistics, Society for Artificial Intelligence and Statistics, Ft. Lauderdale, FL, 1997, 429-437.

Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell Jr. FE, Marks JR, Winchester DP, Bostwick DG. Artificial neural networks improve the accuracy of cancer survival prediction. Cancer 1997;79:857-62.

Burke HB. Evaluating artificial neural networks for medical applications. Proceedings of the 1997 International Conference on Neural Networks, Houston, TX, 1997, 2492 - 2496.

Burke HB. Henson DE. Histologic grade as a prognostic factor in breast carcinoma. Cancer 1997;80:1703-1705.

Burke HB, Hoang A, Iglehart JD, Marks JR. Predicting response to adjuvant and radiation therapy in early stage breast cancer. Cancer, in press.

Burke HB, Henson DE, Taube S. The development and use of tissue banks for prognostic factor research. Submitted for publication.

Burke HB, Henson DE. Screening and early detection. Submitted for publication.

Lundin M, Lundin J, Burke HB, Toikkanen S, Liisa P, Heikki J. The role of time in breast cancer outcome prediction. Submitted for publication.

Presented papers

Burke HB. Creating a Clinical Decision Support System. Grand Rounds, Department of Health Services Research, Mayo Clinic, Rochester MN, October 14, 1996.

Burke HB. Cancer Clinical Decision Support System. Grand Rounds, Department of Pathology, Mayo Clinic, Rochester MN, October 14, 1996.

Burke HB. Medical informatics and clinical decision support systems. Section on Medical Informatics, Stanford University School of Medicine, Stanford CA, November 7 - 8, 1996.

Rosen DB, Burke HB. Applying a gaussian-bernoulli mixture model network to binary and continuous missing data in medicine. Sixth International Workshop on Artificial Intelligence and Statistics, Ft. Lauderdale FL, January 4 -7, 1997.

Burke HB. Creating a cancer clinical decision support system. Specialized Program of Research Excellence in Breast Cancer at Duke University, Duke University Medical Center, February 25-26, 1997.

Burke HB. Reading, using, and performing clinical research (breast cancer). Tutorial: New York State American College of Physicians Associates Annual Meeting, April 5, 1997.

Burke HB. Creating a clinical decision support system for primary care physicians. Primary Care Research Forum, New York Medical College, Valhalla NY, April 23, 1997.

Burke HB. A clinical decision support system for cancer. University of Michigan Cancer Center, Ann Arbor MI, June 9, 1997.

Burke HB. Evaluating artificial neural networks for medical applications. 1997 International Conference on Neural Networks, Houston TX, June 9 - 12, 1997.

Activities Related to this Grant During The Third Funding Year

Sections and Panels

1996 - Present *Chair*, National Institutes of Health, National Cancer Institute Working Group for Research on Prognostic Factors and Systems, Bethesda MD.

1997 - Present *Member*, Screening and Early Detection Advisory Task Force, National Institutes of Health, National Cancer Institute, Bethesda, MD.

Conference positions

1997 *Co-chair*, Special Session on Biomedical Applications, International Congress on Neural Networks, June 11, 1997, Houston Tx.

Artificial Neural Networks Improve the Accuracy of Cancer Survival Prediction

Harry B. Burke, M.D., Ph.D.[1]
Philip H. Goodman, M.D., M.S.[2]
David B. Rosen, Ph.D.[1]
Donald E. Henson, M.D.[3]
John N. Weinstein, M.D., Ph.D.[4]
Frank E. Harrell, Jr., Ph.D.[5]
Jeffrey R. Marks, Ph.D.[6]
David P. Winchester, M.D.[7]
David G. Bostwick, M.D.[8]

[1] Bioinformatics and Health Services Research, Department of Medicine, New York Medical College, Valhalla, New York.

[2] Department of Medicine, University of Nevada School of Medicine, Reno, Nevada.

[3] Division of Cancer Prevention and Control, National Cancer Institute, Bethesda, Maryland.

[4] Division of Cancer Treatment, National Cancer Institute, Bethesda, Maryland.

[5] Department of Health Evaluation Sciences, University of Virginia School of Medicine, Charlottesville, Virginia.

[6] Department of Surgery, Duke University, Durham, North Carolina.

[7] Department of Surgery, Evanston Hospital, Evanston, Illinois; Commission on Cancer, American College of Surgeons, Chicago, Illinois.

[8] Department of Pathology, Mayo Clinic and Mayo Foundation, Rochester, Minnesota.

BACKGROUND. The TNM staging system originated as a response to the need for an accurate, consistent, universal cancer outcome prediction system. Since the TNM staging system was introduced in the 1950s, new prognostic factors have been identified and new methods for integrating prognostic factors have been developed. This study compares the prediction accuracy of the TNM staging system with that of artificial neural network statistical models.

METHODS. For 5-year survival of patients with breast or colorectal carcinoma, the authors compared the TNM staging system's predictive accuracy with that of artificial neural networks (ANN). The area under the receiver operating characteristic curve, as applied to an independent validation data set, was the measure of accuracy.

RESULTS. For the American College of Surgeons' Patient Care Evaluation (PCE) data set, using only the TNM variables (tumor size, number of positive regional lymph nodes, and distant metastasis), the artificial neural network's predictions of the 5-year survival of patients with breast carcinoma were significantly more accurate than those of the TNM staging system (TNM, 0.720; ANN, 0.770; $P < 0.001$). For the National Cancer Institute's Surveillance, Epidemiology, and End Results breast carcinoma data set, using only the TNM variables, the artificial neural network's predictions of 10-year survival were significantly more accurate than those of the TNM staging system (TNM, 0.692; ANN, 0.730; $P < 0.01$). For the PCE colorectal data set, using only the TNM variables, the artificial neural network's predictions of the 5-year survival of patients with colorectal carcinoma were significantly more accurate than those of the TNM staging system (TNM, 0.737; ANN, 0.815; $P < 0.001$). Adding commonly collected demographic and anatomic variables to the TNM variables further increased the accuracy of the artificial neural network's predictions of breast carcinoma survival (0.784) and colorectal carcinoma survival (0.869).

CONCLUSIONS. Artificial neural networks are significantly more accurate than the TNM staging system when both use the TNM prognostic factors alone. New prognostic factors can be added to artificial neural networks to increase prognostic accuracy further. These results are robust across different data sets and cancer sites. *Cancer* **1997; 79:857–62.** © *1997 American Cancer Society.*

KEYWORDS: TNM staging system, artificial neural networks, prognostic factors, breast carcinoma, colorectal carcinoma, survival, outcomes, decision-making, clinical trials, quality assurance.

Presented at the annual meeting of the American Joint Committee on Cancer, Scottsdale, Arizona, January 14, 1995.

Supported in part by research grants from the American Cancer Society (CCG-274), the National Cancer Institute (CA 11606-17), the U.S. Army Medical Research and Development Command Breast Cancer Research Program (DAMD 17-94-J-4383), the Agency for Health Care Policy and Research (HS 06830), and the American Joint Committee on Cancer.

The authors thank John H. Hellier, M.S., for his assistance on this project.

Address for reprints: Harry B. Burke, M.D., Ph.D., Bioinformatics and Health Services Research, Department of Medicine, New York Medical College, Valhalla, NY 10595.

Received May 3, 1996; revision received October 4, 1996; accepted October 15, 1996.

The TNM staging system originated as a response to the need for an accurate, consistent, universal cancer outcome prediction system.[1] Since the TNM staging system was introduced in the 1950s, new prognostic factors have been identified[2,3] and new methods for integrating prognostic factors have been developed.[3] These methods may be capable of (1) providing more accurate predictions than the TNM staging system, using the TNM variables alone (primary tumor size, regional lymph node involvement, and distant metastasis), and (2) further increasing prognostic accuracy by integrating new prognostic factors with the TNM variables. This study compares the cancer specific 5-year survival prediction accuracy for breast and colorectal carcinoma of the TNM staging system with that of artificial neural network statistical models.

METHODS

Data

We used the Commission on Cancer's breast and colorectal carcinoma Patient Care Evaluation (PCE) data sets and the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) breast carcinoma data set.

In October 1992, the American College of Surgeons (ACS) requested cancer information from ACS-accredited hospital tumor registries in the United States. Specifically, they requested the first 25 cases of first-diagnosis breast and colorectal carcinoma seen at each institution in 1983, as well as follow-up information, including deaths, through the date of the request. Variables from this data set used in the breast carcinoma analysis were age, race, payment method, menopausal status, family history, previous biopsy, other cancer, other breast carcinoma, nipple discharge, mammogram, where in the breast the carcinoma occurred, necrosis, histologic grade, estrogen receptor status, progesterone receptor status, number of lymph nodes positive, number of lymph nodes examined, presence or absence of distant metastasis, tumor size, tumor type (in situ, extension to chest wall, or inflammatory), treatment (surgery, chemotherapy, or radiation therapy), and patient outcome (alive or dead). All variables were binary except age, tumor size, number of positive lymph nodes, and number of lymph nodes examined. The PCE data set contained up to 8 years of follow-up information. The analysis end point was breast carcinoma specific 5-year survival. Cases with missing data and those censored before 5 years were excluded. The data set was randomly divided into a training set of 5169 cases, including training and stop-training subsets, and a validation set of 3102 cases.

Variables from the PCE data base used in the colorectal carcinoma analysis were age, race, gender, signs and symptoms (changes in bowel habits, obstruction, jaundice, malaise, occult blood, abdominal pain, pelvic pain, rectal bleeding, or others), diagnostic and extent-of-disease tests (endoscopy, radiography, barium enema, computed tomography scan, biopsy, carcinoembryonic antigen, X-ray, colonoscopy, flexible sigmoidoscopy, intravenous pyelography, liver function tests, biopsy, or other tests), primary site of tumor, level of tumor, histology, grade, number of lymph nodes examined, number of lymph nodes positive, distant metastases, and patient outcome (alive or dead). The end point was 5-year colorectal carcinoma specific survival. After removing cases with missing data and censored patients, the data set was randomly divided into a set of 5007 training cases, including training and stop-training subsets, and a validation set of 3005 cases.

The National Cancer Institute's SEER breast carcinoma data set, for new cases collected from 1977–1982, with 10-year follow-up, was also analyzed. The extent-of-disease variables for the SEER data set were comparable to, but not always identical with, the TNM variables. The end point was breast carcinoma specific 10-year survival. After removing cases with missing data and censored patients, the data set was randomly divided into a set of 3788 training cases, including training and stop-training subsets, and a validation set of 2999 cases.

Models

The TNM staging system used in this analysis was the pathologic system based on the American Joint Committee on Cancer's *Manual for Staging of Cancer*.[1] The TNM staging system's predicted survival for a patient in a particular stage is the average survival of patients in that stage.

In medical research, the most commonly used artificial neural networks (ANN) are multilayer perceptrons that use backpropagation training (Figure 1). Backpropagation consists of fitting the parameters (weights) of the model by a criterion function, usually squared error or maximum likelihood, using a gradient optimization method. In backpropagation artificial neural networks, the error (the difference between the predicted outcome and the true outcome) is propagated back from the output to the connection weights in order to adjust the weights in the direction of minimum error. (For a more detailed description of artificial neural networks, see Burke[4] and Cross.[5]) The artificial neural network employed in this research was composed of three interconnected layers of nodes: an input layer, with each input node corresponding to a patient variable; a hidden layer; and an output layer. All nodes after the input layer sum the inputs to them and use a transfer function (also known as an activa-

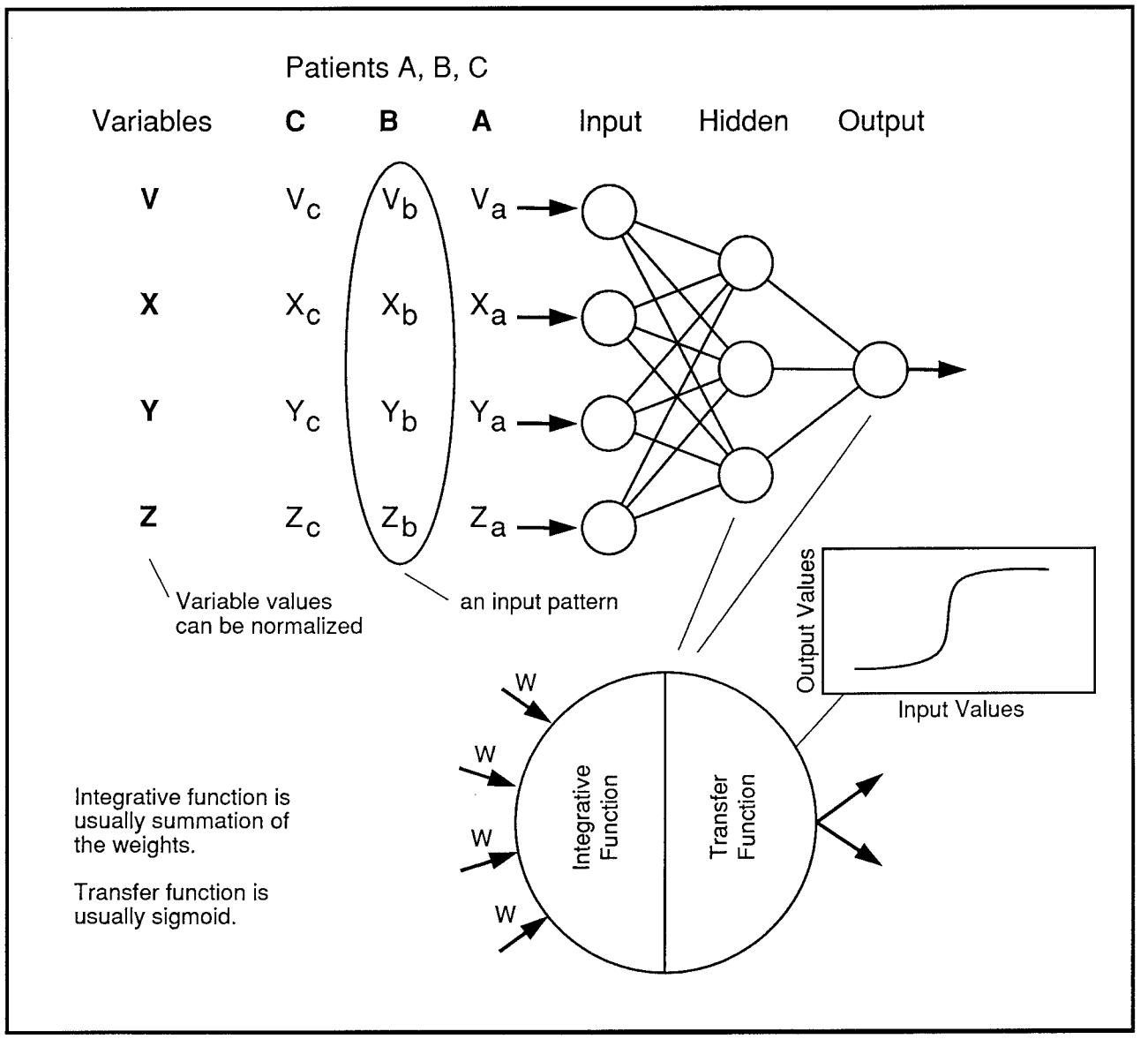

FIGURE 1. Patient A's variable values (Va–Za) are entered into the artificial neural network, followed by patient B, etc. Each variable's input value is multiplied by the weight between the input node for that variable and each hidden layer node it is connected to. All the weighted values going to a hidden layer node are summed at the hidden layer node and go through a sigmoid function before being transferred to the output node. All the weighted values coming into the output node are again summed and put through a sigmoid function. For each patient, the output is a probability from 0–1.0. In training the artificial neural network, the output of each patient is compared with each patient's true outcome. The weights are adjusted so that the next time the patient is presented to the network, the network output is closer to the true outcome.

tion function) to send the information to the adjacent layer nodes. The transfer function is usually a sigmoid function, e.g., the logit. The connections between the nodes have adjustable weights that specify the extent to which the output of one node will be reflected in the activity of the adjacent layer nodes. These weights, along with the connections among the nodes, determine the output of the network.

The mathematical representation of an artificial neural network shown here is equivalent to the graphic model in Figure 1:

$$h_j = f(w^h_{j1}x_1 + w^h_{j2}x_2 + \cdots + w^h_{jn}x_n) \quad (1)$$

$$o_j = g(w^o_1h_1 + w^o_2h_2 + \cdots + w^o_nh_n) \quad (2)$$

where "h_j," in Equation 1 is the output of each of the hidden nodes j, f is a nonlinear transfer function, w^h is the weight from predictor i to hidden node j, and

x_i is an input variable. In Equation 2, o_j is the prediction of the network, g is a nonlinear transfer function, w^o is the weight to the output node, and h is the hidden node output. It should be noted that Equation 2, without the input from Equation 1, is equivalent to logistic regression, where g is the logistic function, w is the beta coefficient, and h is the x covariate.

Specifically, our artificial neural network (NevProp software implementation) used backpropagation training, the maximum likelihood criterion function, and a gradient descent optimization method. The number of input nodes correspond to the number of input variables, the number of hidden layer nodes ranged from three to five, and there was one output mode. Significant differences in the receiver operating characteristic areas between the TNM staging system and the artificial neural network were tested according to the method of Hanley and McNeil.[6] The training data set was divided into training and stop-training subsets. (Training was stopped when accuracy started to decline on the stop-training data subset.) All analyses employed the same training and validation data sets, and all results were based on the one-time use of the validation data sets.

Accuracy

There are three components to predictive accuracy: the amount and quality of the data, the predictive power of the prognostic factors, and the prognostic method's ability to capture the power of the prognostic factors. This study focused on the third component.

The measure of comparative accuracy is the trapezoidal approximation to the area under the receiver operating characteristic curve.[7] The area under this curve is a nonparametric measure of discrimination. While squared error summarizes how close each patient's prediction is to the true outcome, the receiver operating characteristic area measures the relative goodness of the set of predictions as a whole by comparing the predicted probability of each patient with that of all pairs of patients. This area is calculated using the predictive scores of each algorithm in order to compare their average accuracy in predicting outcome. The receiver operating characteristic area is independent of both the prior probability of each outcome and the threshold cutoff for categorization, and its computation requires only that the algorithm produce an ordinally-scaled relative predictive score. In terms of mortality, the receiver operating characteristic area estimates the probability that the algorithm will assign a higher mortality score to the patient who died than to the patient who lived. The receiver operating characteristic area varies from 0 to 1. When the prognostic score is unrelated to survival, the score is 0.5, indicating chance accuracy. The farther the score is from 0.5, the better, on average, the prediction model is at predicting which of the two patients will be alive.

TABLE 1
Comparison of the TNM Staging System with the Artificial Neural Network

Data sets	TNM staging system	Artificial neural network
PCE breast CA, TNM variables alone	0.720	0.770[a]
PCE breast CA, TNM and added variables	0.720	0.784[a]
SEER breast CA, TNM variables alone	0.692	0.730[b]
PCE colorectal CA, TNM variables alone	0.737	0.815[a]
PCE colorectal CA, TNM and added variables	0.737	0.869[a]

PCE: Patient Care Evaluation (Commission on Cancer); SEER: Surveillance, Epidemiology, and End Results (National Cancer Institute).
[a] $P < 0.001$.
[b] $P < 0.01$.

RESULTS

A comparison of the accuracy of the TNM staging system and the artificial neural network is shown in Table 1. For the PCE breast carcinoma data set, using only the TNM variables (tumor size, number of positive regional lymph nodes, and distant metastasis), the artificial neural network's predictions of breast carcinoma specific 5-year survival were significantly more accurate than those of the TNM staging system (TNM 0.720; vs. ANN, 0.770, $P < 0.001$). Since the TNM staging system is, by definition, limited to the TNM variables, additional variables do not improve the TNM staging system's predictive accuracy. However, adding commonly collected demographic and anatomic variables to the TNM variables further increased the accuracy of the artificial neural network (to 0.784).

We were able to test whether the artificial neural network's significant improvement in predictive accuracy was generalizable across data sets. For the National Cancer Institute's 1977–1982 SEER breast carcinoma data set, using only the TNM variables, the artificial neural network's predictions of 10-year survival were significantly more accurate than those of the TNM staging system (TNM 0.692 vs. ANN 0.730, $P < 0.01$).

We were able to test whether the artificial neural network's significant improvement in predictive accuracy was generalizable across cancer sites. For the PCE colorectal data set, using only the TNM variables, the artificial neural network's predictions of 5-year colorectal carcinoma specific survival were significantly more accurate than those of the TNM staging system (TNM 0.737 vs. ANN 0.815, $P < 0.001$). Adding commonly collected demographic and anatomic variables

to the TNM variables further increased the accuracy of the artificial neural network (0.869).

To clarify the clinical importance of the observed increases in accuracy, we changed the area under the curve (A_z) scale to a −1 to +1 scale, i.e., [2(A_z − 0.5)]. On this scale, 0 was chance and 1.0 was perfect prediction. By this measure, the TNM staging system's accuracy was 44% greater than chance for breast carcinoma specific 5-year survival predictions. Placing the TNM variables in the artificial neural network increased predictive accuracy to 54%, and adding variables that individually had little prognostic value to the artificial neural network further increased prognostic accuracy to 57% greater than chance prediction. Corresponding increases in predictive accuracy specific to colorectal carcinoma were as follows: 47% for the TNM staging system increased to 63% when the TNM variables were placed in the artificial neural network, and that increased to 74% when several commonly collected variables were added to the artificial neural network.

DISCUSSION

The TNM staging system is only moderately accurate in its breast and colorectal carcinoma specific 5-year survival predictions. The significant superiority in predictive accuracy that the artificial neural network showed when compared with the TNM staging system across data sets and cancer sites suggests that it is able to improve our ability to predict the survival of cancer patients. In addition, artificial neural networks can be expanded to include any number of prognostic factors. They can accommodate continuous variables and they can provide presurgery and postsurgery treatment predictions.

Artificial neural networks are a class of nonlinear regression and discrimination statistical methods. They are of proven value in many areas of medicine.[8–19] They do not require a priori information regarding the phenomenon, and they make no distributional assumptions. When the appropriate method is used to avoid overfitting (i.e., loss of generalization by fitting the patterns to the test data too precisely), artificial neural networks are usually at least as accurate as classical statistical models, and, depending on the complexity of the phenomena, they can be much more accurate. In predicting 5-year breast carcinoma specific survival, they have been shown to be more accurate than logistic regression, classification and regression trees (CART; pruned or shrunk), and principal components analysis.[20]

The improvement in prognostic ability made possible by artificial neural networks may be clinically important for therapy, clinical trials, patient information, and quality assurance. In decision-making regarding therapy, it may allow the efficient separation of patients with a poor prognosis (who require therapy) from patients with an excellent prognosis (who require little or no therapy), and it may predict who will respond to a particular therapy. In clinical trials, it may decrease interpatient variability. This would allow for the creation of more homogenous patient populations for clinical trials, resulting in smaller clinical trial patient populations, less expensive trials, and the ability to detect treatment effects that would be undetectable in more heterogeneous study populations. With regard to patient information, it may give patients a clearer understanding of the time course of their disease. Finally, for assessment and quality assurance, it may provide a better severity of illness adjustment.

REFERENCES

1. Beahrs OH, Henson DE, Hutter RVP, Kennedy BJ, editors. American Joint Committee on Cancer. Manual for staging of cancer. 4th edition. Philadelphia: JB Lippincott, 1992.
2. Burke HB, Hutter RVP, Henson DE. Breast carcinoma. In: P Hermanek, MK Gospadoriwicz, DE Henson, RVP Hutter, LH Sobin, editors. UICC prognostic factors in cancer. Berlin: Springer-Verlag, 1995: 165–76.
3. Burke HB, Henson DE. Criteria for prognostic factors and for an enhanced prognostic system. *Cancer* 1993;72:3131–5.
4. Burke HB. Artificial neural networks for cancer research: outcome prediction. *Semin Surg Oncol* 1994;10:1–7.
5. Cross SS, Harrison RF, Kennedy RL. Introduction to neural networks. *Lancet* 1995;346:1075–9.
6. Hanley JA, McNeil BJ. The meaning of the use of the area under the receiver operating characteristic (ROC) curve. *Radiology* 1982;143:29–36.
7. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic. *J Math Psy* 1975;12;387–415.
8. Baxt WG. Application of artificial neural networks to clinical medicine. *Lancet* 1995;346:1135–8.
9. Dybowski R, Gant V. Artificial neural networks in pathology and medical laboratories. *Lancet* 1995;346:1203–7.
10. Westenskow DR, Orr JA, Simon FH. Intelligent alarms reduce anesthesiologist's response time to critical faults. *Anesthesiology* 1992;77:1074–9.
11. Tourassi GD, Floyd CE, Sostman HD, Coleman RE. Acute pulmonary embolism: artificial neural network approach for diagnosis. *Radiology* 1993;189:555–8.
12. Leong PH, Jabri MA. MATIC: an intracardiac tachycardia classification system. *Pacing Clin Electrophysiol* 1992; 15:1317–31.
13. Gabor AJ, Seyal M. Automated interictal EEG spike detection using artificial neural networks. *Electroencephalogr Clin Neurophysiol* 1992;83:271–80.
14. Goldberg V, Manduca A, Ewert DL. Improvement in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence. *Med Phys* 1992; 19:1275–81.
15. O'Leary TJ, Mikel UV, Becker RL. Computer-assisted image interpretation: use of a neural network to differentiate tubular carcinoma from sclerosing adenosis. *Mod Pathol* 1992;5:402–5.
16. Dawson AE, Austin RE, Weinberg DS. Nuclear grading of breast carcinoma by image analysis. *J Clin Pathol* 1991; 95(Suppl):S29–S37.

17. Wu Y, Giger ML, Doi K, Vyborny CJ, Schmidt RA, Metz CE. Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer. *Radiology* 1993;187:81–7.
18. Astin ML, Wilding P. Application of neural networks to the interpretation of laboratory data in cancer diagnosis. *Clin Chem* 1992;38:34–8.
19. von Osdol W, Myers TG, Paull KD, Kohn KW, Weinstein JN. Use of the Kohonen self-organizing map to study the mechanisms of action of chemotherapeutic agents. *J Natl Cancer Inst* 1994;86:1853–9.
20. Burke HB, Rosen DB, Goodman PH. Comparing the prediction accuracy of artificial neural networks and other statistical models for breast cancer survival. In: Tesauro G, Touretzky DS, Leen TK, editors. Advances in neural information processing systems 7. Cambridge, MA: MIT Press, 1995: 1063–7.

EDITORIAL

Counterpoint

Histologic Grade as a Prognostic Factor in Breast Carcinoma

Harry B. Burke, M.D., Ph.D.[1]
Donald Earl Henson, M.D.[2]

[1] Bioinformatics and Health Services Research, Department of Medicine, New York Medical College, Valhalla, New York.

[2] Division of Cancer Prevention, National Cancer Institute, National Institutes of Health, Bethesda, Maryland.

In this issue of *Cancer,* Dr. Roberti reviews the role of histologic grade in the prognosis of breast carcinoma and wonders why, because it is available, it has not been widely used in predicting outcome.[1] The position of this editorial is that there must be some fundamental reason, after 100 years of progress on histologic grade, that confusion persists regarding its prognostic value.

The systematic use of morphologic variation at the cellular level of analysis as a prognostic factor in cancer has been fraught with controversy. Currently, there is no universally agreed on set of necessary and sufficient conditions for the definition of histologic grade in breast carcinoma. There has been uncertainty regarding the identification of what variation was important, how the variation should be organized, and whether it should be integrated into a staging or index system.

An additional issue is that grading system criteria have been selected based on their ability to create subgroups of patients using histologic distinctions to produce significant differences in outcome. There are two problems with this approach. First, there are many possible criteria that can create significant differences between subgroups and there is no analytic method for finding the best criteria.[2] Second, statistical significance is not necessarily accuracy. Significance is the chance that two or more distributions of variables, as represented by their parameter estimates, for example, means and variances, are really the same. Accuracy assesses the strength of association between two or more variables.[3,4] In general, accuracy quantifies how good a variable is at predicting another variable. Specifically, we are interested in the strength of association between grade and survival, i.e., how good is grade at predicting survival.

Fundamentally, grade remains controversial because it confounds two types of time. One type is how long the tumor has been growing and the other is how rapidly it has been growing. A "high grade" tumor could be an indolent tumor that grew for a long time prior to discovery and will continue to be slow growing; alternatively, it could be an aggressive tumor of recent origin that will continue to be rapidly growing. Because one can never know when a tumor originated, it may not be possible on histologic grounds to separate

Supported in part by a research grant from the U.S. Army Medical Research and Development Command Breast Cancer Research Program (DAMD 17-94-J-4383).

See reply to counterpoint on pages 1706–7 and referenced original article on pages 1708–16, this issue.

Address for reprints: Donald Earl Henson, M.D., Division of Cancer Prevention, National Cancer Institute, Bethesda, MD 20892.

Received May 9, 1997; accepted May 15, 1997.

a slowly growing tumor from a rapidly growing tumor. In other words, one cannot always distinguish how long the tumor has been growing from how fast it has been growing. The extent to which time ambiguity exists in grade is the extent to which grade's prediction variance will increase and consequently the extent to which its prediction accuracy will decrease. This limits grade's independent prognostic value and its ability to add significant prognostic value when placed in a system that includes other time-related factors such as tumor size.

The mechanical theory of cancer, a view espoused by Halsted,[5] assumes that cancer spreads from the primary tumor to the regional lymph nodes and then to distant sites of the body. This view is the basis of the TNM staging system. For the mechanical theory, the primary purpose of a prognostic system is to capture the spread of the cancer because cancer spread is believed to be the best indicator of outcome. The three elements of the TNM staging system (local tumor, regional lymph node, and distant metastasis[6]) are believed to reflect directly the spread of cancer, i.e., the extent of disease. Grade is not one of the TNM variables because it does not fit into this mechanical epistemology; it does not directly reflect the spread of the cancer. However, even if grade could have been subsumed within the mechanical theory of cancer, it would not have replaced tumor size in breast carcinoma. Using the Surveillance, Epidemiology, and End Results data of the National Cancer Institute for 1983–1987 and the area under the receiver operating characteristic curve as the measure of accuracy (Az), we found the Az for grade alone to be .634 and the Az for tumor size alone to be .737 ($P < .05$) for 5-year survival. Furthermore, grade does not add prognostic accuracy to tumor size; the Az for tumor size and grade combined was .749, which was not significant when compared with tumor size alone. In addition, grade could not have been added to the TNM staging system because the system is a bin model comprised of five levels of tumor characteristics (T), four levels of regional lymph node involvement (N), and two levels of distant metastasis (M).[7] Adding the 4 levels of grade to the 40 bins of the TNM ($5T \times 4N \times 2M$) would have created 160 bins and made it too complex to be useful.[7]

What is the future of grade as a prognostic factor in breast carcinoma? If we no longer accept the mechanical theory of cancer spread, grade becomes a possible prognostic factor. In addition, because the TNM staging system is not very accurate[8] new computer-based prognostic systems are being developed.[7] Computer-based prognostic systems are more accurate in predicting outcome and they do not have a limitation on the number of variables that can be used. Can grade be an independent prognostic factor in a computer-based system or can grade substitute for another more difficult to assess factor such as lymph node status?

We evaluated the ability of grade to predict 5-year breast carcinoma survival using data from the National Cancer Institute's SEER program.[9] The data were collected between 1983–1987 and the patients were followed for at least 5 years. The variables were tumor size, local extent of disease, lymph node status, and histologic grade. The criteria used to determine grade were neither standardized nor explicitly reported. The data set did not include cases with metastatic disease because grade is infrequently reported in these patients. Only 14,704 of the 48,643 cases were graded (30%). All analyses without grade were performed on the full data set of 48,643 cases. An analysis using the subset of graded cases favors grade because it is almost certainly the case that the variance of grade would increase if all the cases were graded. The area under the receiver operating characteristic curve was the measure of prediction accuracy. We used the logistic regression statistical method to create our models (SAS Institute, Cary, NC) and all results were performed on the test data set.

The predictive accuracy of tumor size, local tumor extent, and lymph status was .794. Adding histologic grade slightly increased the Az to .797, but this was not significant. Therefore, in a statistical model with traditional prognostic factors, grade does not add prognostic accuracy.

Can histologic grade substitute for a factor that is becoming difficult to evaluate (e.g., lymph node status). To answer this question, we created a logistic regression model in which grade was the predictor and lymph node metastasis (detected vs. not detected) was the outcome. This addressed the issue of how well grade can take the place of lymph node status as a prognostic factor (in other words, to what extent does their prognostic information overlap?). If their predictions completely overlap, then the observed Az would be 1.0; if there was no overlap, then the observed Az would be .5. Again using the SEER data set, we found an Az of .589, which indicated that there was very little predictive overlap. Therefore, grade is not an effective surrogate for nodal status.

If grade is to be a useful prognostic factor in the future it must improve predictive accuracy for women with small tumors and few involved lymph nodes when used in predictive models that include the new molecular genetic prognostic factors. The data set from Duke University, kindly provided by Dr. Jeffrey Marks, includes patients with early stage breast carcinoma. These data were described in a previous arti-

cle.[10] Briefly, all patients were pathologic TNM Stage I or early Stage II. Early Stage II included all TNM Stage II patients except those with five or more positive lymph nodes. The variables were age, race, tumor size, positive lymph nodes, TNM lymph node status, nuclear grade, histologic grade, p53, c-*erb* B-2 (HER-2/*neu*), estrogen receptor status (ER) and progesterone receptor status (PR), vascular invasion, adjuvant therapy (tamoxifen, chemotherapy), and radiation therapy. Patients who underwent a lumpectomy received radiation therapy. Patients who underwent a modified radical mastectomy did not receive radiation therapy. There were 229 cases, 226 of which had complete data for all variables except ER and PR status. Because many individual patient ER and PR values were missing, both variables were removed from the data set. The 5-year survival rate was 70%. The logistic regression statistical method was used to create the models and a prediction endpoint of 5-year overall survival.

Neither histologic grade nor nuclear grade added any predictive power to the new molecular genetic prognostic factors in the logistic regression model. The predictive accuracy for all factors excluding histologic and nuclear grade was .733; when histologic grade was added the Az was .738 (not significant), when nuclear grade was added the Az was .736 (not significant), and when both were added the Az was .740 (not significant).

Overall, the accuracy of the Duke University logistic regression models was lower than the SEER logistic regression models because outcome prediction for early stage breast carcinoma was more difficult than outcome prediction for early and late stage breast carcinoma. In the Duke data set, the TNM staging system performed at chance level when predicting the outcome of women with early stage breast carcinoma, the Az was .567.[11]

Histologic grade alone has modest prognostic value. However, grade does not significantly increase the predictive accuracy of computer-based prognostic systems, either in data sets that represent all stages of breast carcinoma and contain traditional predictive factors or in data sets that represent early stage breast carcinoma and contain the new molecular genetic prognostic factors.

REFERENCES

1. Roberti NE. The role of histologic grading in the prognosis of patients with carcinoma of the breast: is this a neglected opportunity? *Cancer* 1997;80:1708–66.
2. Burke HB. Integrating multiple clinical tests to increase predictive accuracy. In: Hanausek M, Walaszek, Z, editors. Methods in molecular biology: tumor marker protocols. Mahwah, NJ: Lawrence Erlbaum Associates, 1997. In press.
3. Hays WL. Statistics. 5th edition. Fort Worth, TX: Harcourt Brace, 1991:335–8.
4. Burke HB. Evaluating artificial neural networks for medical applications. Proceedings of the 1997 International Congress of Neural Networks, 1997.
5. Halsted WS. The results of radical operations for the cure of carcinoma of the breast. *Ann Surg* 1907;46:1–19.
6. Beahrs OH, Henson DE, Hutter RVP, Kennedy BJ, editors. Manual for staging of cancer. 4th edition. Philadelphia: J.B. Lippincott, 1992.
7. Burke HB, Henson DE. Criteria for prognostic factors and for an enhanced prognostic system. *Cancer* 1993;72:3131–5.
8. Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell FE Jr., et al. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer* 1997;79:868–73.
9. Shambaugh EM, Ries LG, Young JL Jr. Extent of disease: new 4-digit schemes, codes and coding instructions. Bethesda, MD: National Institutes of Health, National Cancer Institute, Biometry Branch, 1984.
10. Marks JR, Wu K, Berry D, Bandarenko N, Kerns BJ, Iglehart JD. Overexpression of p53 and HER-2/neu proteins as prognostic markers in early stage breast cancer. *Ann Surg* 1994;219:332–41.
11. Burke HB, Hoang A, Iglehart JD, Marks JR. Predicting response to adjuvant and radiation therapy in early stage breast cancer. *Cancer*. In press.

Predicting Response to Adjuvant and Radiation Therapy in Patients with Early Stage Breast Carcinoma

Harry B. Burke, M.D., Ph.D.[1]
Albert Hoang, Ph.D.[1]
J. Dirk Iglehart, M.D.[2]
Jeffrey R. Marks, Ph.D.[2]

[1] Department of Medicine, New York Medical College, Valhalla, New York.

[2] Departments of Surgery, Pathology, and Cell Biology, Duke University, Durham, North Carolina.

Supported in part by a research grant from the U.S. Army Medical Research and Development Command Breast Cancer Research Program (DAMD 17-94-J-4383) and by the Duke Specialized Program of Research (SPORE) of the National Cancer Institute.

Address for reprints: Harry B. Burke, M.D., Ph.D., Department of Medicine, New York Medical College, Valhalla, NY 10595.

Received March 14, 1997; revision received July 31, 1997; accepted September 8, 1997.

BACKGROUND. Screening and surveillance is increasing the detection of early stage breast carcinoma. The ability to predict accurately the response to adjuvant therapy (chemotherapy or tamoxifen therapy) or postlumpectomy radiation therapy in these patients can be vital to their survival, because this prediction determines the best postsurgical therapy for each patient.

METHODS. This study evaluated data from 226 patients with TNM Stage I and early Stage II breast carcinoma and included the variables p53 and c-*erb*B-2 (HER-2/*neu*). The area under the receiver operating characteristic curve (Az) was the measure of predictive accuracy. The prediction endpoints were 5- and 10-year overall survival.

RESULTS. For Stage I and early Stage II patients, the 5- and 10-year predictive accuracy of the TNM staging system were at chance level, i.e., no better than flipping a coin. Both the 5- and 10-year artificial neural networks (ANNs) were very accurate—significantly more so than the TNM staging system (Az 5-year survival, TNM = 0.567, ANN = 0.758; $P < 0.001$; Az 10-year survival, TNM = 0.508, ANN = 0.894; $P < 0.0001$). For patients not receiving postsurgical therapy and for either chemotherapy or tamoxifen therapy, the ANNs containing p53 and c-*erb*B-2 and the number of positive lymph nodes were accurate predictors of survival (Az 5-year survival, 0.781, 0.789, and 0.720, respectively).

CONCLUSIONS. The molecular genetic variables p53 and c-*erb*B-2 and the number of positive lymph nodes are powerful predictors of survival, and using ANN statistical models is a powerful method for predicting responses to adjuvant therapy or radiation therapy in patients with breast carcinoma. ANNs with molecular genetic prognostic factors may improve therapy selection for women with early stage breast carcinoma. ***Cancer* 1998;82:874–7.** © *1998 American Cancer Society.*

KEYWORDS: TNM staging system, artificial neural networks, prognostic factors, breast carcinoma, tamoxifen therapy, chemotherapy, radiation therapy, outcomes, c-*erb*B-2, p53.

Screening and surveillance is increasing the prevalence of early stage breast carcinoma. The ability to predict accurately the responses to adjuvant therapy (chemotherapy or tamoxifen therapy) or postlumpectomy radiation therapy in these patients can be vital to their survival, because this prediction determines the best postsurgical therapy for each patient. The pathologic TNM staging system is the current cancer prognostic system. Its predictions are based on three variables: 1) location, size, and depth of tumor; 2) existence and location of involved lymph nodes; and 3) existence of distant metastases.[1] We have shown that artificial neural networks (ANNs)

are more accurate at predicting survival than the TNM staging system for all stages of breast carcinoma.[2] It is not known how accurate the TNM staging system is in predicting the survival of patients with early stage breast carcinoma. It is also not known whether ANNs with molecular genetic prognostic factors, i.e., p53 and c-*erb*B-2 (HER-2/*neu*), can improve prognostic accuracy in early stage breast carcinoma across postsurgical therapies and for specific therapies. This article compares the survival prediction accuracy of the TNM staging system with ANN models across all postsurgical therapies. In addition, it presents a method for properly assessing putative therapy-dependent prognostic factors and examines the accuracy of ANNs in terms of specific therapies. Because the TNM staging system does not predict response to adjuvant or radiation therapy, it is not included in the individual therapy analyses.

METHODS

Data

These data were described in detail in a previous article.[3] Briefly, all patients were pathologic TNM Stage I or early Stage II. Early stage breast carcinoma includes Stage I and limited Stage II. Limited Stage II included all the TNM Stage II patients except those with five or more positive lymph nodes. The variables were age, race, tumor size, lymph nodes positive, lymph node stage, nuclear grade, histologic grade, p53, c-*erb*B-2, estrogen receptor (ER) and progesterone receptor (PR) status, vascular invasion, adjuvant therapy (tamoxifen or chemotherapy), and radiation therapy. Patients who underwent a lumpectomy received radiation therapy. Patients who underwent a modified radical mastectomy did not receive radiation therapy. There were 229 cases, of which 226 had complete data for all variables except ER and PR status. Because of the number of cases missing, both ER and PR were removed from the data set. The survival rate was 70%. The prediction endpoints were 5-and 10-year overall survival.

Accuracy

The area under the receiver operating characteristic curve (Az) is a measure of prediction accuracy.[4] It can be used to assess and compare the adequacy of statistical models. Az can be directly calculated by Somer's D,[5] or it can be approximated by its trapezoidal area.[6] The area under the curve is a nonparametric measure of discrimination. It is independent of both the prior probability of each outcome and the threshold cutoff for category. Its computation requires only that the prediction method produce an ordinally scaled relative predictive score. In terms of mortality, the receiver operating characteristic area estimates the probability that the prediction method will assign a higher mortality score to the patient who died than to the patient who lived. The receiver operating characteristic area varies from 0 to 1. When the predictions are unrelated to survival, the score is 0.5, indicating chance accuracy. The farther the score is from 0.5, the better, on average, the prediction method is for predicting which of the two patients will be alive.

TABLE 1
Comparison of the Accuracy of the TNM Staging System and Artificial Neural Networks in Predicting the 5- and 10-year Survival of Patients with Early Stage Breast Carcinoma

Model	5-yr survival Az (SE)[a]	10-year survival Az (SE)[b]
TNM	0.567 (0.046)	0.508 (0.053)
ANN	0.758 (0.042)	0.894 (0.034)

ANN: artificial neural network; Az: area under the receiver operating characteristic curve; SE: standard error.
[a] TNM vs. ANN 5-year survival, $P < 0.001$.
[b] TNM vs. ANN 10-year survival, $P < 0.0001$.

Statistical Models

ANN models have been described in detail elsewhere.[2] Briefly, the three-layer backpropagation ANN was composed of an input layer, a hidden layer, and an output layer. Each layer of an ANN was composed of nodes. The number of input nodes was equal to the number of variables. The hidden layer was composed of three nodes. There was one output node. All the variables were entered into the three-layer ANN model. The two-layer ANN was identical to the three-layer network, except that it did not possess a hidden layer. After a sensitivity analysis to reduce the number of input variables to the three with the highest predictive accuracy, the three selected variables, namely, the number of positive lymph nodes, p53, and c-*erb*B-2, were entered into the two-layer ANN. Both the two- and three-layer ANNs employed the maximum likelihood loss function and weight decay. Model accuracy estimates and standard errors were calculated by the bootstrap resampling method.[7]

RESULTS

The predictive accuracies of the TNM staging system and the three-layer ANN models are shown in Table 1. For Stage I and early Stage II patients, the 5- and 10-year prediction accuracy of the TNM staging sys-

TABLE 2
Artificial Neural Network Accuracy in Predicting 5-Year Survival for Each Theory Combination

Strata	T	C	R	No. of cases	5-yr survival Az (SE)
1	–	–	–	48	0.781 (0.091)
2	–	–	+	23	
3	–	+	–	53	0.789 (0.049)
4	–	+	+	19	
5	+	–	–	43	0.720 (0.072)
6	+	–	+	7	
7	+	+	–	14	
8	+	+	+	19	

T: tamoxifen; C: chemotherapy; R: radiation; Az: area under the receiver operating characteristic curve; SE: standard error; +: patient received the therapy; –: patient did not receive the therapy.

tem was at chance level, i.e., no better than flipping a coin. Both the 5- and 10- year ANNs were very accurate and significantly more accurate than the TNM staging system (Az 5- year survival, TNM = 0.567, ANN = 0.758, $P < 0.001$; Az 10-year survival, TNM = 0.508, ANN = 0.894, $P < 0.0001$).

The evaluation of therapy-dependent prognostic factors requires the mutually exclusive and exhaustive partitioning of the adjuvant therapies and radiation therapy. Because the numbers of patients and outcomes were small in this and in the subsequent analyses, three variables (number of positive lymph nodes, p53, and c-*erb*B-2) and two-layer ANNs with a 5-year survival endpoint were employed. The stratification by postsurgical therapy into eight bins is shown in Table 2.

There was a no-therapy bin (Stratum 1) and there were bins representing all combinations of the three postsurgical therapies, i.e., tamoxifen, chemotherapy, and radiation therapy (Strata 2–8). Only Stratum 1 (no adjuvant therapy), Stratum 3 (only chemotherapy), and Stratum 5 (only tamoxifen) contained enough patients for analysis. The ANNs for these three strata were accurate predictors of survival (Az 5-year survival, 0.781, 0.789, and 0.720, respectively).

The number of cases in each bin could be increased by stratifying by therapy regardless of whether a patient received another therapy. This was not a mutually exclusive and exhaustive partitioning of the therapy variables. Thus, the results must be viewed as an approximation, because the variables were not being treated as purely therapy-dependent prognostic factors. Table 3 shows the accuracy of the ANN for each of the three therapies. With larger numbers in each strata, it is clear that the ANNs that contained the three variables lymph nodes positive, p53, and c-*erb*B-2 were excellent predictors of response to adjuvant therapy (Az 5-year survival, tamoxifen = 0.855, chemotherapy = 0.782, radiation = 0.861).

DISCUSSION

We have demonstrated that ANNs that contain p53 and c-*erb*B-2 are significantly more accurate than the TNM staging system at predicting 5- and 10-year survival in women with early stage breast carcinoma. We have also demonstrated that the molecular genetic variables p53 and c-*erb*B-2 and the number of positive lymph nodes can be used to accurately predict responses to surgery, chemotherapy, and tamoxifen therapy.

An understanding of therapy-dependent prognostic factors, and why there must be a mutually exclusive and exhaustive partitioning of the therapies prior to the assessment of therapy-dependent prognostic factors, requires a description of the types and functions of prognostic factors. Prognostic factor types are defined in terms of their function. There are three prognostic factor functions and therefore three types of prognostic factors: natural history, therapy-dependent, and posttherapy.[8] Natural history prognostic factors predict the course of the disease if no effective therapy exists or if an effective therapy is not administered. For example, clinically palpable lymph nodes may be a natural history prognostic factor. Therapy-dependent prognostic factors predict, prior to the patient's receiving the therapy, a change in the course of the disease caused by a change in the patient's

TABLE 3
Artificial Neural Network Accuracy in Predicting 5-Year Survival for Each Therapy

Treatment group	No. of cases	5-yr survival Az (SE)
Tamoxifen	83	0.855 (0.052)
Chemotherapy	105	0.782 (0.055)
Radiation	68	0.861 (0.047)

Az: area under the receiver operating characteristic curve; SE: standard error.

condition due to receipt of an effective therapy. For example, ER status may predict response to tamoxifen. Posttherapy prognostic factors predict, after the patient has received the therapy, whether there has been a change in the course of the disease due to the intervention. For example, the number of positive lymph nodes on axillary dissection may predict whether the patient will respond to the primary surgery. Posttherapy prognostic factors are important because we do not want to wait any longer than necessary to administer a second-line therapy to patients who do not respond to the primary therapy. All three prognostic factors are relative to therapy. For each therapy in a succession of therapies (for example, if a therapy is given and the patient does not respond to that therapy and another therapy is contemplated), all three types of prognostic factors can be analyzed.

Within the context of the small sample size of this study, the molecular genetic variables p53 and c-*erb*B-2 are powerful therapy-dependent prognostic factors for early stage breast carcinoma, and ANN models are an efficient statistical method for capturing their predictive power.

REFERENCES

1. Beahrs OH, Henson DE, Hutter RVP, Kennedy BJ, editors.. Manual for staging of cancer. American Joint Committee on Cancer. 4th edition. Philadelphia: J. B. Lippincott, 1992.
2. Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell Jr. FE, et al. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer* 1997;79: 857–62.
3. Marks JR, Wu K, Berry D, Bandarenko N, Kerns BJ, Iglehart JD. Overexpression of p53 and HER-2/*neu* proteins as prognostic markers in early stage breast cancer. *Ann Surg* 1994; 219:332–41.
4. Swets JA. Signal detection theory and ROC analysis in psychology and diagnostics: collected papers. Mahwah, NJ: Lawrence Erlbaum Associates, 1996.
5. Somers RH. A new asymmetric measure of association for ordinal variables. *Am Sociological Rev* 1962;27:799–811.
6. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic. *J Math Psy* 1975;12:387–415.
7. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall, 1993:45–57.
8. Burke HB. Increasing the power of surrogate endpoint biomarkers: the aggregation of predictive factors. *J Cell Biochem* 1994;19S:278–82.